Dummy Variables

A **dummy variable** or **binary variable** is a variable that takes on a value of 0 or 1 as an indicator that the observation has some kind of characteristic. Common examples:

- Sex (female): FEMALE=1 if individual in the observation is female, equal to 0 otherwise
- Race (White): WHITE=1 if individual in the observation is white/Caucasian, equal to 0 otherwise
- Urban vs Rural: URBAN=1 if individual in the observation lives in an urban area, equal to 0 otherwise
- College graduate: COLGRAD=1 if individual in the observation has a four-year college degree, equal to 0 otherwise

It is common to use dummy variables as explanatory variables in regression models, if binary categorical variables are likely to influence the outcome variable.

1. Example: Factors Affecting Monthly Earnings

Let us examine a data set that explores the relationship between total monthly earnings (`MonthlyEarnings`) and a number of variables on an interval scale (i.e. numeric quantities) that may influence monthly earnings including including each person's IQ (`IQ`), a measure of knowledge of their job (`Knowledge`), years of education (`YearsEdu`), and years experience (`YearsExperience`), years at current job (`Tenure`).

The data set also includes dummy variables that may explain monthly earnings, including whether or not the person is black / African American (`Black`), whether or not the person lives in a Southern U.S. state (`South`), and whether or not the person lives in an urban area (`Urban`).

The code below downloads a CSV file that includes data on the above variables from 1980 for 935 individuals and assigns it to a data set that we name `wages`.

```
wages <- read.csv("http://murraylax.org/datasets/wage2.csv");
```

The following call to `lm()` estimates a multiple regression predicting monthly earnings based on the eight explanatory variables given above, which includes three dummy variables. The next call to `summary()` displays some summary statistics for the estimated regression.

```
lmwages <- lm(MonthlyEarnings
              ~ IQ + Knowledge + YearsEdu + YearsExperience + Tenure
              + Black + South + Urban,
              data = wages)
summary(lmwages)
```

```
##
## Call:
## lm(formula = MonthlyEarnings ~ IQ + Knowledge + YearsEdu + YearsExperience +
##     Tenure + Black + South + Urban, data = wages)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -874.42 -229.18  -40.25  181.26 2163.02
##
```

```
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -451.0098   121.3752  -3.716 0.000215 ***
## IQ                 2.5966     0.9963   2.606 0.009301 **
## Knowledge          6.5545     1.8142   3.613 0.000319 ***
## YearsEdu          47.6530     7.1378   6.676 4.22e-11 ***
## YearsExperience   12.4833     3.1746   3.932 9.04e-05 ***
## Tenure             6.2910     2.4049   2.616 0.009043 **
## Black           -110.6660    39.2222  -2.822 0.004882 **
## South            -50.8222    25.7903  -1.971 0.049068 *
## Urban            155.4316    26.4621   5.874 5.94e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 356.7 on 926 degrees of freedom
## Multiple R-squared:  0.2285, Adjusted R-squared:  0.2219
## F-statistic: 34.29 on 8 and 926 DF,  p-value: < 2.2e-16
```

The p-values in the right-most column reveal that all of the coefficients are statistically significantly different from zero at the 5% significance level. We have statistical evidence that all of these variables influence monthly earnings.

The coefficient on `Black` is equal to -110.67. This means that even after accounting for the effects of all the other explanatory variables in the model (includes educational attainment, experience, location, knowledge, and IQ), black / African American people earn on average \$110.67 less per month than non-black people.

The coefficient on `South` is -50.82. Accounting for the impact of all the variables in the model, people that live in Southern United States earn on average \$50.82 less per month than others.

The coefficient on `Urban` is 155.43. Accounting for the impact of all the variables in the model, people that live in urban areas earn \$155.43 more per month, which probably reflects a higher cost of living.

We can compute confidence intervals for these effects with the following call to `confint()`

```r
confint(lmwages, parm=c("Black", "South", "Urban"), level = 0.95)
```

```
##            2.5 %      97.5 %
## Black -187.6407 -33.6913263
## South -101.4365  -0.2079364
## Urban  103.4989 207.3642822
```

2. Dummy Interactions with Numeric Explanatory Variables

We found that black people have lower monthly earnings on average than non-black people. In our regression equation, this implies that the *intercept* is lower for black people than non-black people. We can also test whether a dummy variable affects the *slope* multiplying other variables.

For example, are there differences in the returns to education for black versus non-black people? To answer this, we include an *interaction effect* between `Black` and `YearsEdu`:

```r
lmwages <- lm(MonthlyEarnings
              ~ IQ + Knowledge + YearsEdu + YearsExperience + Tenure
              + Black + South + Urban + Black*YearsEdu,
              data = wages)
summary(lmwages)
```

```
##
## Call:
## lm(formula = MonthlyEarnings ~ IQ + Knowledge + YearsEdu + YearsExperience +
##     Tenure + Black + South + Urban + Black * YearsEdu, data = wages)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -871.77 -223.35  -39.15  183.60 2166.96
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -484.8569   122.7181  -3.951 8.38e-05 ***
## IQ                 2.5965     0.9951   2.609 0.009224 **
## Knowledge          6.6834     1.8135   3.685 0.000242 ***
## YearsEdu          50.0652     7.2573   6.899 9.73e-12 ***
## YearsExperience   12.0943     3.1784   3.805 0.000151 ***
## Tenure             6.3322     2.4022   2.636 0.008528 **
## Black            328.4032   249.9481   1.314 0.189211
## South            -48.6125    25.7902  -1.885 0.059753 .
## Urban            155.1421    26.4318   5.870 6.09e-09 ***
## YearsEdu:Black   -35.0262    19.6929  -1.779 0.075630 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 356.3 on 925 degrees of freedom
## Multiple R-squared:  0.2312, Adjusted R-squared:  0.2237
## F-statistic:  30.9 on 9 and 925 DF,  p-value: < 2.2e-16
```

We see here that when accounting for an interaction effect between race and education, the coefficient on the `Black` dummy variable becomes insignificant, but the coefficient on the interaction term is negative and significant at the 10% level. The coefficient on the interaction term equal to -35.03 means the slope on education is 35.03 less when $Black = 1$.

The coefficient on the interaction term is interpreted as the *additional* marginal effect of the numeric variable for the group associated with the dummy variable equal to 1. For this example:

- The marginal effect on monthly earnings for non-black people for an additional year of education is equal to \$50.07 (i.e. when $Black = 0$).
- The marginal effect on monthly earnings for black people for an additional year of education is equal to \$50.07 - \$35.03 = \$15.02 (i.e. when $Black = 1$).
- Said another way, the marginal effect on monthly earnings for an additional year of education is \$35.03 less for black people than non-black people.

3. Interacting Dummy Variables with Each Other

Let us interact two of the dummy variables to understand this interpretation and motivation. In the call to `lm()` below, we use our baseline model and interact `South` and `Urban`:

```
lmwages <- lm(MonthlyEarnings
              ~ IQ + Knowledge + YearsEdu + YearsExperience + Tenure
              + Black + South + Urban + South*Urban,
              data = wages)
summary(lmwages)
```

```
##
## Call:
## lm(formula = MonthlyEarnings ~ IQ + Knowledge + YearsEdu + YearsExperience +
##     Tenure + Black + South + Urban + South * Urban, data = wages)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -885.94 -228.09  -36.76  173.16 2153.62
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -516.8840   124.9159  -4.138 3.83e-05 ***
## IQ                 2.7472     0.9968   2.756 0.005964 **
## Knowledge          6.6968     1.8118   3.696 0.000232 ***
## YearsEdu          48.1580     7.1275   6.757 2.50e-11 ***
## YearsExperience   12.9375     3.1753   4.074 5.01e-05 ***
## Tenure             6.1817     2.4007   2.575 0.010178 *
## Black           -109.0280    39.1521  -2.785 0.005467 **
## South             30.3594    45.5537   0.666 0.505288
## Urban            200.1871    33.5683   5.964 3.51e-09 ***
## South:Urban     -116.3504    53.8671  -2.160 0.031033 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 356 on 925 degrees of freedom
## Multiple R-squared:  0.2324, Adjusted R-squared:  0.2249
## F-statistic: 31.12 on 9 and 925 DF,  p-value: < 2.2e-16
```

To interpret the meaning of the coefficient on `South`, `Urban`, and `South*Urban`, we will ignore (hold constant) all the terms in the regression equation that do not include one of these variables.

3.1 Difference between Urban and Rural Workers in the North/East/West

Workers in the North / East / and West U.S. have $South = 0$. Here $South = 0$, $(South\ x\ Urban) = 0$, so neither the coefficient on the interaction nor the coefficient on `South` come into play.

The coefficient for b_{Urban} implies that *in the Non-Southern U.S.*, urban workers earn on average \$200.19 more in monthly earnings than rural workers.

3.2 Difference between Urban and Rural Workers in the South

When focusing on workers in the South, $South = 1$ and the interaction term comes into play.

- Impact for urban workers in the south $= b_{South}(1) + b_{Urban}(1) + b_{Urban*South}(1)$
- Impact for rural workers in the south $= b_{South}(1) + b_{Urban}(0) + b_{Urban*South}(0)$
- Difference $= b_{Urban} + b_{Urban*South} = 200.19 - 116.35 = \$83.84$

In the Southern U.S. states, urban workers on average earn \$83.84 more in monthly earnings than rural workers.

3.3 Difference between Southern and North/East/West Monthly Earnings for ***Urban*** Workers

- Impact for Southern urban workers = $b_{South}(1) + b_{Urban}(1) + b_{Urban*South}(1)$
- Impact for Non-Southern urban workers = $b_{South}(0) + b_{Urban}(1) + b_{Urban*South}(0)$
- Difference = $b_{South} + b_{Urban*South} = 30.36 - 116.35 =$ -\$85.99

For urban workers, workers in the South earn \$85.99 less in monthly earnings than workers outside the South.

3.4 Difference between Southern and North/East/West Monthly Earnings for ***Rural*** Workers

Rural workers have $Urban = 0$ and so the interaction term $Urban\ x\ South = 0$, so we can ignore both of those coefficients. The coefficient for b_{South} implies that Southern rural workers earn on average \$30.36\$ *more* per month than Non-Southern rural workers.

4 Three-Way Interactions and Higher!

What?! Things aren't complicated enough for you?! Do at your own peril!

I have seen people include higher order interaction effects like `South * Urban * Black * YearsEdu` in their regressions. It has never been obvious to me that they understood what their results meant.